

Building Tools for Image-Based Electronic Editions

Ionut E. Iacob and Alex Dekhtyar

The EPT serves to organize the raw materials of digital scholarship – digital image and text files – and, using specialized encoding, builds these materials into a usable electronic edition. This edition will include a wide variety of editorial information, including organizational description of both physical (books, folios, lines) and semantic (sentences, words), glossarial and metrical description, and description of the condition of the physical object, notably how that condition interacts with the text on the page. For this reason, it is vital for a successful IBEE that the EPT enable the editor to create links between the images and text.

The eXtensible Markup Language (XML) is preferred by the humanities computing community as data support for electronic text encoding, most notably through Guidelines of the Text Encoding Initiative. Although XML does not well capture complex text structures (its strict hierarchical organization severely limits its usefulness in describing, for example, both physical and textual organization in a single file), its relative simplicity recommends it over more powerful but complex representations. Moreover, XML is well supported by software processing tools, from databases, parsers and editors (supporting syntax coloring and on-the-fly validation) to query engines and XML transformations. Many good XML editors are available at no, or very low, cost, which makes XML an even more attractive choice for humanities text encoding.

Building an electronic edition is a tedious enterprise. The editor using traditional XML software must encode editorial information while remaining mindful of XML syntax and the limits imposed by its use. A misplaced tag can keep an XML file from validating, and often an editor will have to choose between encoding different aspects of the manuscript text or risk overlapping markup (for example, the physical organization of a folio – the lines as they appear on the page – may conflict with the sentence structure of the text). Things become more complicated when images are involved. The editor has to keep track of images and record relationships between text and image, not just relating entire folios to the text on that folio, but identifying corresponding regions of text and image. The unfortunate result of this process is that as the complexity of the encoding increases, the editor must concentrate on the syntax of encoding rather than on the details of the text of the manuscript or edition. Our goal was to design tools that allow the editor to concentrate on the act of editing, rather than focus on issues of XML syntax and validity.

As James Clark points out, there are two main classes of XML editors: *text editors* and *structural editors*. The key difference between these two kinds of editors is the way markup is introduced. Structural editors focus on data-centric encoding, and the editing process begins with markup. The human editor adds content to an encoding template, in a manner similar to entering items in a database. This is in contrast to text editors, which focus on document-centric editing and begin with the textual content (PCDATA). The editor inserts markup into

(or deletes it from) the content one tag at a time. The text editor approach is much preferable for humanities editing in general and image-based encoding specifically, as it gives the human editor control over exactly what markup is entered where in the text. This control is important for image-based editing, as it facilitates the recording of image-text relationships by allowing the human editor to select specific sections of text and, with the right software support, relate that text to the corresponding sections of image. Another issue that arises in document-centric encoding is that the XML document may not be valid during the editing process: the order in which the editor introduces the markup in the text may depend not on the requirements of the DTD, but rather on the *modus operandi* of the human editor (which in turn depends on the semantics of the features to be encoded).

Thus, an image-based XML editor has to have the following features:

- Hide the XML syntax if requested. The focus of the human editor should be on text semantics and how images and text are connected. Instead of displaying the complete XML, show where markup exists by highlighting the relevant text in the display. The editor may at times wish to examine the XML encoding. In that case, the XML editor should provide a system for filtering out unwanted markup, showing only those elements that the editor wishes to see.
- Allow text markup by enabling the editor to select the range of content to be marked up and the tag (and attribute values) to be inserted. Among tag attributes, at least one is dedicated to link text and corresponding image or image region.
- Provide support for the editor to connect the markup with the corresponding manuscript image and a specific region in the image. While the editor selects the related areas, the information for mapping the image to the text should be saved automatically by the software – the editor should not have to concern himself with creating image maps or noting image coordinates.
- Assure document well-formedness and provide support for (partial) validation in such a way that it is transparent to the human editor. Imposing validity constraints for update operations might be too prohibitive in text encoding applications: not every update operation (or a set of consecutive update operations) yields a valid document. The software takes further update decisions based on the current status of a document. At the same time, it is important to be able to verify at each moment of time that the current XML fragment is "on track", i.e., that the human editor has not committed any structural error while introducing the markup (in which case markup deletion is required). We call this *potential validation* and we designed and implemented an algorithm for checking potential validity of document-centric XML documents.
- Provide support for searching for both text and structure, and for searching the encoding of image features described in the XML markup. There are three main types of searches that the editor can perform in an IBEE. First, the text search, through which the editor can search for a string of characters in the edition content. Second the structural search –

this information describes how various text and image features are interrelated (words in certain lines or sentences, holes on the folio in the middle of sentences, etc.). And finally, image feature searches. Given a specified region on the image, the software will find all encoded features related to that region or, conversely, will find all image regions corresponding to a given text range or descriptor (for example, find all image regions with corresponding damage markup).

In this paper, we will describe how we designed and built the EPT and its individual components to incorporate those elements that we found most important for image-based, document-centric editing.

References:

- T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler, eds. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. 6 October 2000. .
- M. S. Brown and W. S. Seales. "The Digital Atheneum: New Approaches for Preserving, Restoring, and Analyzing Damaged Manuscripts." Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM Press, 2001. 437-443.
- M. S. Brown, W. B. Seales, K. Kiernan, and J. Griffioen. "3D Acquisition and Restoration of Medieval Manuscripts." Communications of the ACM: Special Issue on Digital Libraries. May 2001
- J. Clark. *Incremental XML Parsing and Validation in a Text Editor*, December 2003. Presentation at XML 2003, Philadelphia.
- D. Hayes. "Glossing Damaged Manuscripts: an Example from Ælfric's Lives of Saints." Digital Resources for the Humanities (DRH01). University of London, London, UK. 10 July 2001.
- K. Kiernan, A. Dekhtyar, J. Jaromczyk, D. C. Porter, and I. E. Iacob. "Edition Production Technolog (EPT) and the ARCHway Project." DigiCULT.Info, 8 (August 2004), 36-38.
- W. B. Seales, J. Griffioen, K. Kiernan, C. J. Yuan, and L. Cantara. "The Digital Atheneum: New Technologies for Restoring and Preserving Old Documents." Computers in Libraries 20:2 (February 2000), 26-30.
- C.M. Sperberg-McQueen, L. Burnard, eds. Guidelines for Text Encoding and Interchange (P4), The TEI Consortium, 2001.
- C. J. Yuan and W. B. Seales. "Guided Linking: Efficiently Making Image-to-Transcript Correspondence." Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM Press, 2001, 471.